A MATHEMATICAL PERSPECTIVE ON TRANSFORMERS

BORJAN GESHKOVSKI, CYRIL LETROUIT, YURY POLYANSKIY, AND PHILIPPE RIGOLLET

arXiv:2312.10794v5

10/06/25

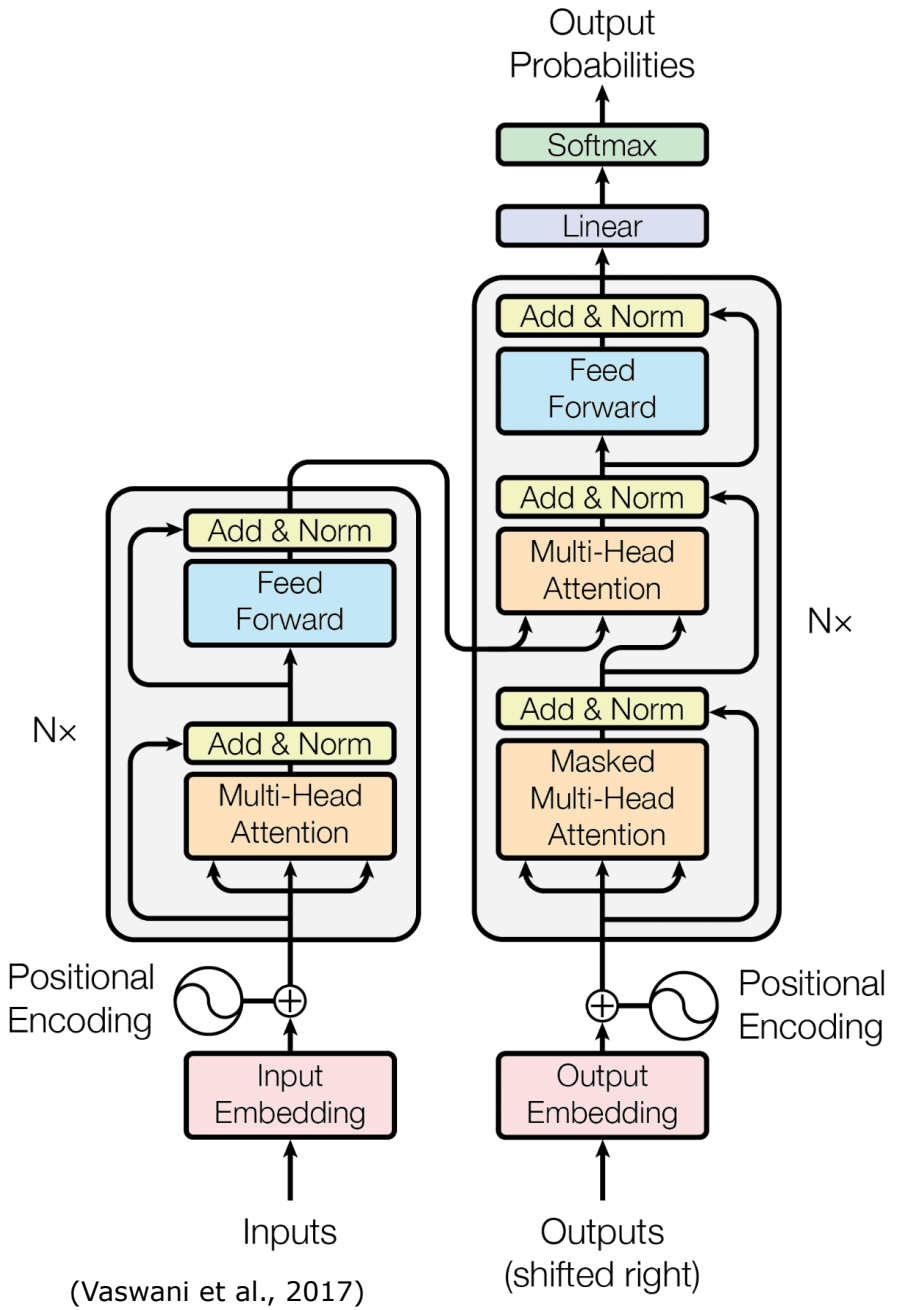
STAT 946 – Mathematics of Deep Learning

Presentation by Marty Mukherjee,
Department of Applied Mathematics

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

A BRIEF INTRO TO TRANSFORMERS

- An encoder and/or decoder architecture that consists of scaled dot-product attention mechanism.
  - Good representation of compatibility
  - Fast and interpretable computation
  - Parallelizable evaluation across all queries (can leverage GPUs)
  - Scaled dot-products for stable softmax gradients in high dimensions (prevents large magnitudes)



(Vaswani et al., 2017)

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

A BRIEF INTRO TO TRANSFORMERS

Input + Positional Encoding: $\boldsymbol{x}(0) \in \mathbb{R}^{n \times d}$

- n tokens, each with dimension d

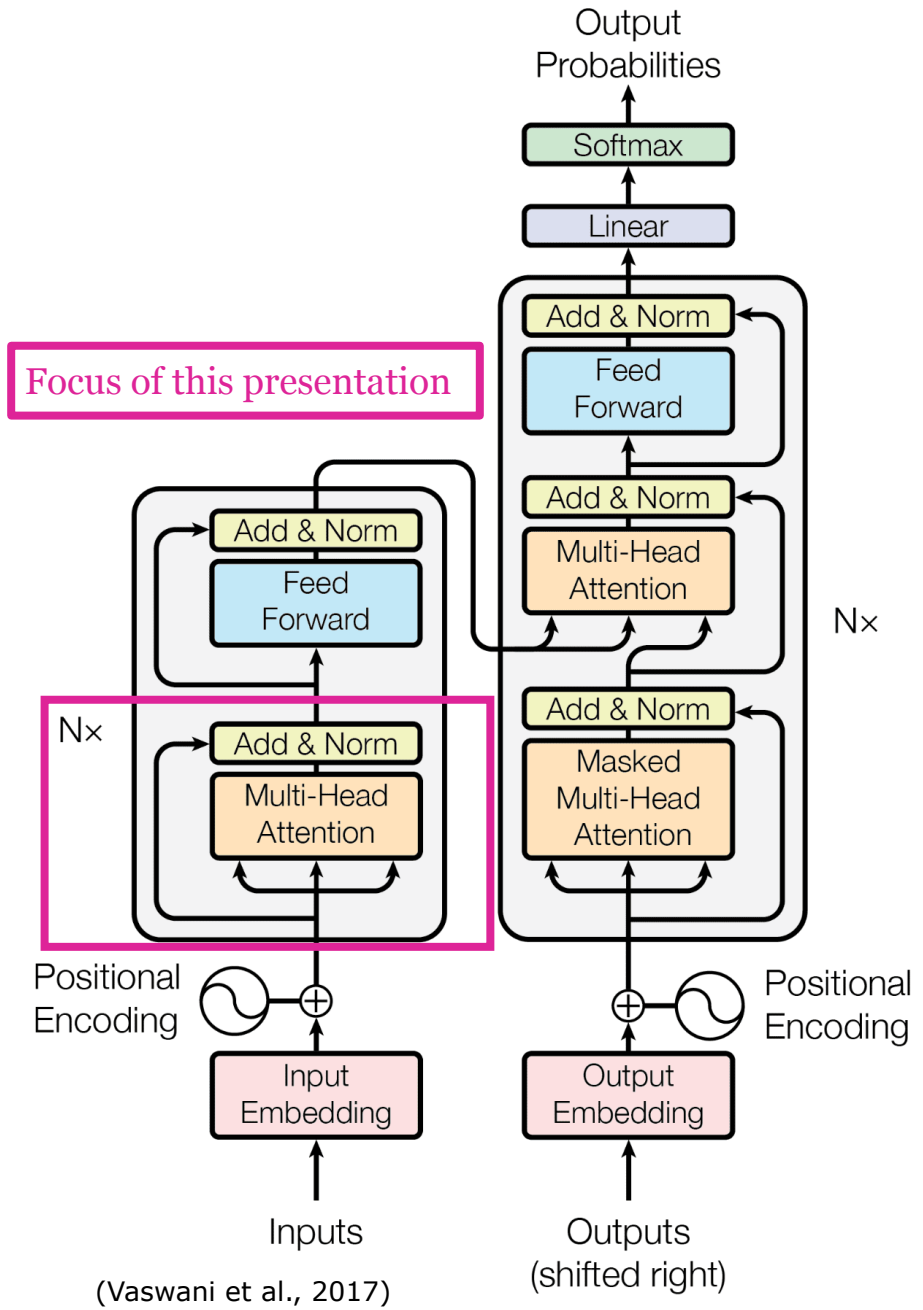
Attention (Single-Head for this presentation, denoted $SAtt$):

- Key: $K(k) \in \mathbb{R}^{n \times l}$, Query: $\mathrm{Q}(k) \in \mathbb{R}^{n \times l}$, Value: $\mathrm{V}(k) \in \mathbb{R}^{n \times l}$
- $SAtt(\mathbf{x}(k); K(k), Q(k), V(k)) = Softmax\left(\beta^{-1}(Q(k)\mathbf{x}(k)^T(K(k)\mathbf{x}(k))\right) V(k)\mathbf{x}(k)$

Goal: Find the "alignment" between keys and queries to scale values

Update:

$$\boldsymbol{x}(k+1) = \frac{x(k) + SAtt\left(\boldsymbol{x}(k); K(k), Q(k), V(k)\right)}{\left\| x(k) + SAtt\left(\boldsymbol{x}(k); K(k), Q(k), V(k)\right) \right\|}$$

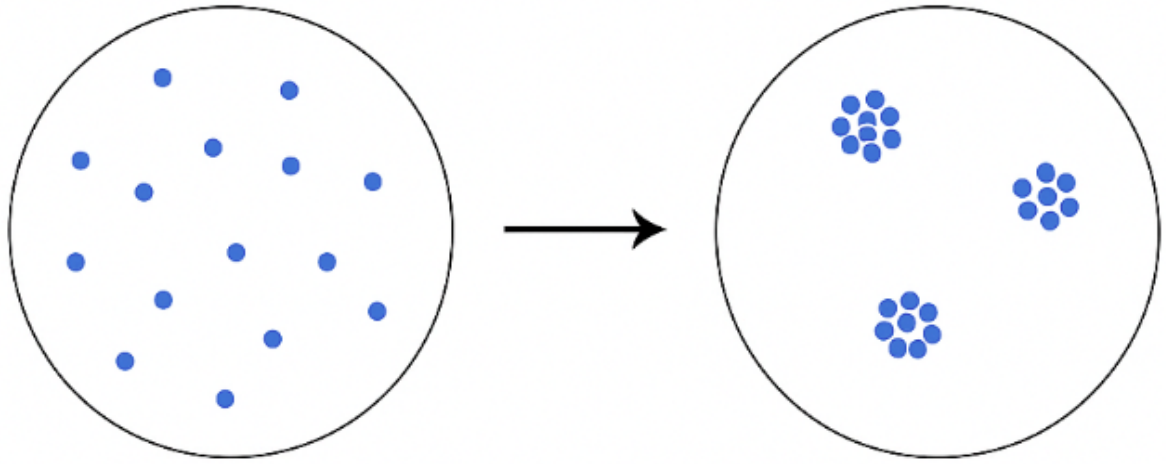


(Vaswani et al., 2017)

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Motivation

- Transformers, like residual neural networks, can be modeled in continuous-time as an ordinary differential equation (ODE).
- Key idea: View transformers as interacting particle systems, where each particle is a token.
- Key observation: Particles tend to “cluster” under these dynamics – limiting distribution is a point mass???



Clustering phenomenon

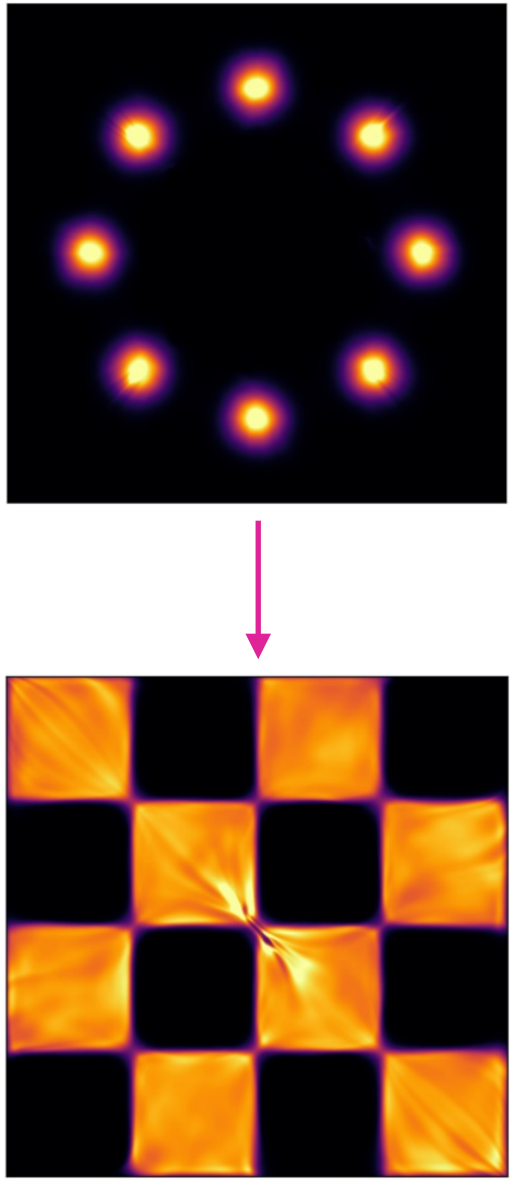
Outline of the Presentation

- Background on optimal transport
  - Continuity equation
  - Wasserstein distance -> Wasserstein gradient
- Mathematical framework to study transformers
  - Probability flows on spheres (self-attention and layer-normalization)
  - Wasserstein gradient flow – convergence in distribution
- Clustering
  - A single cluster for large temperature
  - A single cluster for small temperature

BACKGROUND ON OPTIMAL TRANSPORT

Monge and Kantarovich Problem

- Introduced by Monge in 1781, the optimal transport (OT) problem is concerned with transferring mass from one distribution to another in a way as to minimize an expected cost function $c(\cdot, \cdot)$.
- Let X and Y be two Polish spaces, and let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$.
  Goal: Find a measurable transport map $T : X \times Y$ such that $T_{\#}\mu = \nu$
  and T minimizes $\inf_{T:T_{\#}\mu=\nu} \int_X c(x, T(x))\mu(dx)$
- Kantarovich proposed a relaxation by considering joint distributions $\Pi(\mu, \nu)$: $\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\pi(dx, dy)$



Lai et al. "FP-Diffusion: Improving Score-based Diffusion Models by Enforcing the Underlying Score Fokker-Planck Equation". ICML 2023.

Monge (1781). "Théorie des déblais et des remblais". De l'Imprimerie Royale

Variational Optimal Transport

- If the cost function is a **distance** metric, then we denote the minimum cost $W_p^p(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(x,y)\sim\pi}[d(x,y)^p]$ as the p-**Wasserstein** distance.
- Brenier (1991) proved the existence of a **unique convex potential** $\varphi(x_t)$ such that the vector field $b(x_t) = \nabla\varphi(x_t)$ that solves the OT problem in continuous time.
- Let us consider the ODE:

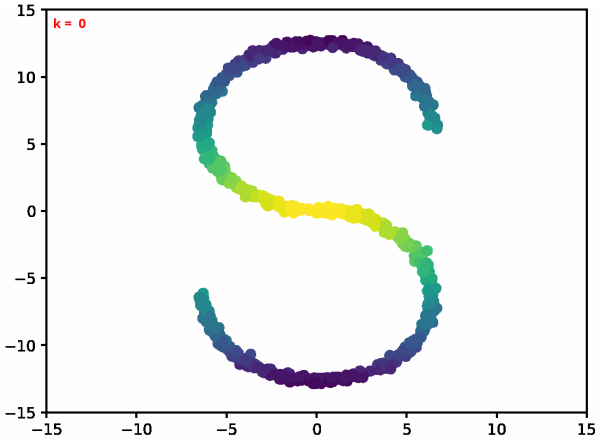
$$\dot{x}_t = b(x_t), x_0 \sim \mu$$

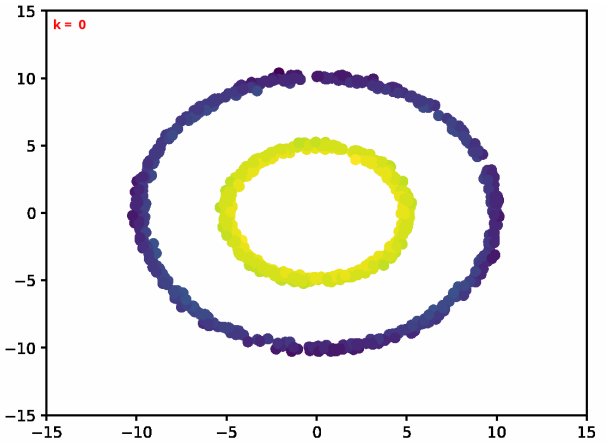
- This ODE evolves a distribution, modelled by the **continuity equation**

$$\partial_t p_t = -\nabla\cdot(p_t b_t), p_0 = \mu, p_1 = \nu$$

- Benamou and Brenier (2000) proved that the 2-Wasserstein distance is equal to the average kinetic energy:

$$W_2^2(\mu,\nu) = \min_{b \text{ s.t. } p_0=\mu, p_1=\nu} \int_0^1 \int p_t(x_t)\|b(x_t)\|^2 dx_t dt$$





Benamou, J.-D. and Y. Brenier (2000). "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". Numerische Mathematik, vol. 84, pp. 375–393.
Brenier, Y. (1991). "Polar factorization and monotone rearrangement of vector-valued functions". Communications on Pure and Applied Mathematics, vol. 44, no. 4, pp. 375–417.

De Bortoli et al. "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". NeurIPS 2021.

Towards understanding Wasserstein gradient - Rethinking Gradients?

- Gradients are simply defined by their underlying inner products.

$$\frac{d}{dt}V(\mathbf{x}(t))\,|_{t=0} = \left\langle \nabla V(\mathbf{x}(0)), \frac{d}{dt}\mathbf{x}(t)\,|_{t=0} \right\rangle$$

- How about function spaces (e.g. L_2)?

$$\frac{d}{dt}f(x_t)\,|_{t=0} = \left\langle \nabla_{L_2} f(x_0), \partial_t x_t\,|_{t=0} \right\rangle_{L_2}, \text{ where } \langle f, g\rangle = \int f(z)g(z)dz$$

- Analogy to Wasserstein spaces - Needs an inner product!!!

Wasserstein Distance as an Inner Product

$$W_2^2(\mu,\nu) = \min_{\varphi \text{ s.t. } p_0=\mu, p_1=\nu} \int_0^1 \int p_t(x_t) \|\nabla \varphi(x_t)\|^2 dx_t dt$$

- Analogy to weighted L_2 norm:

$$W_2^2(\mu,\nu) = \min_{p_t \text{ s.t. } p_0=\mu, p_1=\nu} \int_0^1 \|\partial_t p_t\|_{p_t}^2 dt$$

- (Which is strikingly similar to the geodesic distance on a Riemannian manifold!)

$$\text{where } \|\partial_t p_t\|_{p_t}^2 = \min_{\varphi_t} \|\nabla \varphi(\mathbf{x})\|_{L_2(p_t)}^2 \text{ s.t. } \partial_t p_t = -\nabla \cdot (p_t \nabla \varphi)$$

- We can now define the following Wasserstein inner product on two functions h_1 and h_2 with $\int h_1 = \int h_2 = 0$:

$$\langle h_1, h_2 \rangle_p = \int \langle \nabla \varphi_1, \nabla \varphi_2 \rangle p(d\mathbf{x}) = \langle \nabla \varphi_1, \nabla \varphi_2 \rangle_{L_2(p)}, \text{ where } -\nabla \cdot (p \nabla \varphi_i) = h_i$$

Grand Finale - Wasserstein Gradient

- Use the definition of derivative
- $\frac{d}{dt} f(p_t)\,|_{t=0} = \left\langle \nabla_{W_2} f(p_0), \partial_t p_t\,|_{t=0} \right\rangle_{p_t}$
- Assume f is L_2-differentiable

$$\frac{d}{dt} f(p_t)\,|_{t=0} = \left\langle \nabla_{L_2} f(p_0), \partial_t p_t\,|_{t=0} \right\rangle_{L_2} = -\left\langle \nabla_{L_2} f(p_0), \nabla \cdot (p_0 \nabla \varphi) \right\rangle_{L_2}$$

$$= \left\langle \nabla \nabla_{L_2} f(p_0), p_0 \nabla \varphi \right\rangle_{L_2} = \left\langle \nabla \nabla_{L_2} f(p_0), \nabla \varphi \right\rangle_{L_2(p_0)}$$

- Comparing to the Wasserstein inner product, we conclude

$$\nabla_{W_2} f(p) = -\nabla \cdot (p \nabla \nabla_{L_2} f(p))$$

Otto, F. (2001). "The geometry of dissipative evolution equations: the porous medium equation". Communications in Partial Differential Equations, vol. 26, no. 1-2, pp. 101–174.

Digesting Wasserstein Gradient Flows

- Revisit continuity equation. Suppose a density evolves by

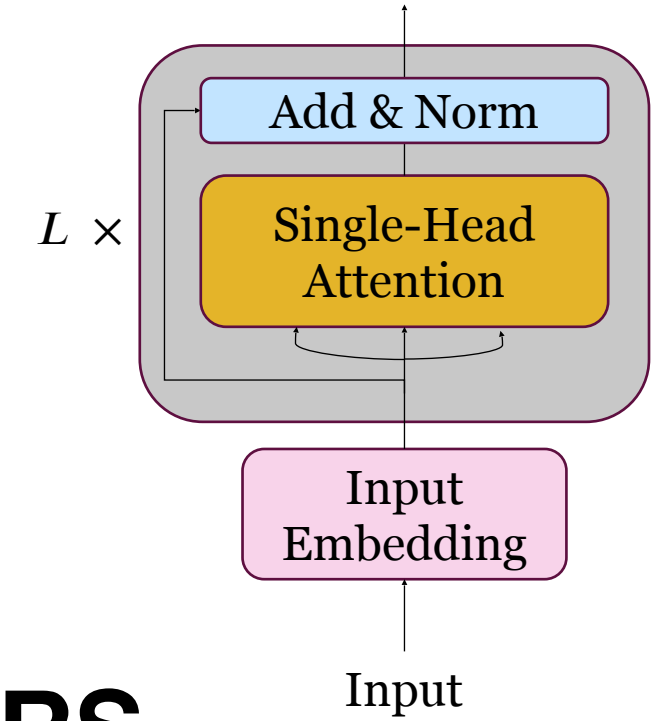
$$\partial_t p_t = -\nabla_{W_2} f(p_t) = \nabla \cdot (p \nabla \nabla_{L_2} f(p))$$

- Then a particle drawn from p_t evolves by the ODE

$$\dot{\mathbf{x}}_t = \nabla \nabla_{L_2} f(p_t(\mathbf{x}_t))$$

- Benefits of Wasserstein gradient flow:
  - Strong stability guarantees
  - Guaranteed conservation of mass

- Example: If $f(p) = \int u(p(\mathbf{x}))d\mathbf{x}$, then
  - $\nabla_{L_2} f(p) = u'(p)$
  - $\nabla_{W_2} f(p) = -\nabla \cdot (p u''(p) \nabla p)$
- If $f(p) = \int u(\mathbf{x})p(\mathbf{x})d\mathbf{x}$,
  - $\nabla_{L_2} f(p) = u(p)$
  - $\nabla_{W_2} f(p) = -\nabla \cdot (p \nabla u)$



MODELLING OF TRANSFORMERS

- Probability flows on spheres (self-attention and layer-normalization)
- Wasserstein gradient flow – convergence in distribution

Inspiration from Residual Neural Networks (ResNets)

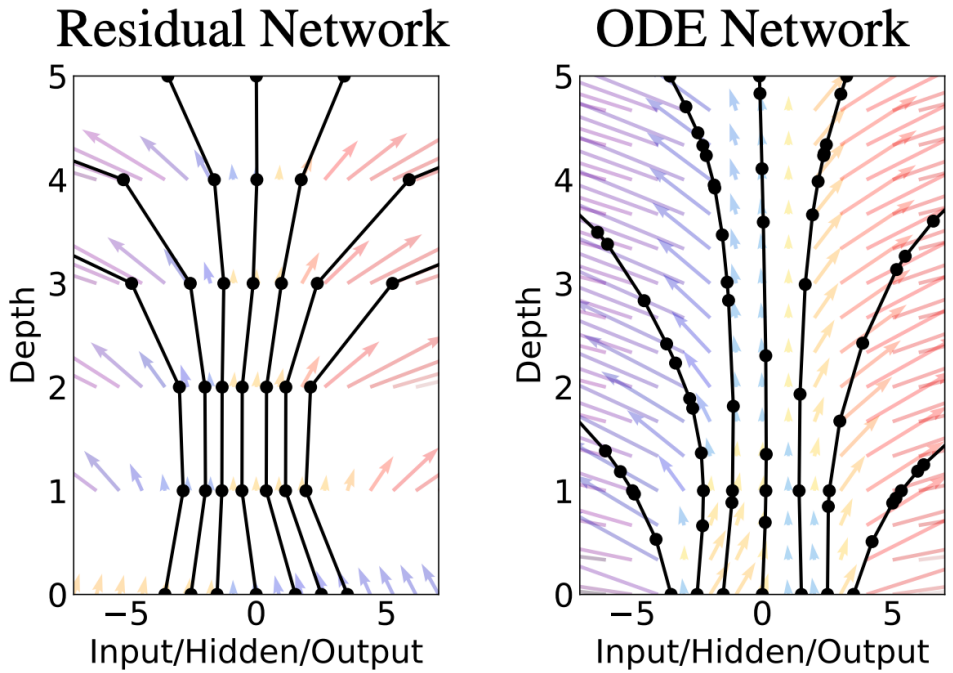
- ResNets approximate a function f at $x \in \mathbb{R}^d$ through a sequence of layers with skip connections

$$\boldsymbol{x}(k+1) = \boldsymbol{x}(k) + w(k)\sigma\big(a(k)\boldsymbol{x}(k) + b(k)\big),\ k = 0,\ldots,L-1,\ \boldsymbol{x}(0) = \boldsymbol{x}$$

- The layer can be interpreted naturally in continuous-time

$$\dot{\boldsymbol{x}}(t) = w(t)\sigma\big(a(t)\boldsymbol{x}(t) + b(t)\big),\ t \in [0,T],\quad \boldsymbol{x}(0) = \boldsymbol{x}$$

- These are called *neural ODEs*.



The role of Softmax – Understanding Self-Attention

$$SAtt(\mathbf{x}(k)) = Softmax(\beta(Q(k)\mathbf{x}(k)) \cdot (K(k)\mathbf{x}(k)))\ V(k)\mathbf{x}(k)$$

- Expanding $Softmax$

$$\left[SAtt\left(\boldsymbol{x}(k)\right)\right]_i = \sum_{j=1}^{n} \frac{\exp\left(\beta\left\langle Q(k)\boldsymbol{x}_i(k), K(k)\boldsymbol{x}_j(k)\right\rangle\right)}{Z_{\beta,i}(k)} V(k)\boldsymbol{x}_j(k),$$

Where $Z_{\beta,i}(k) = \sum_{j=1}^{n} \exp\left(\beta\left\langle Q(k)\boldsymbol{x}_i(k), K(k)\boldsymbol{x}_j(k)\right\rangle\right)$ is the normalization term, and

$A_{ij}(k) = \exp\left(\beta\left\langle Q(k)\boldsymbol{x}_i(k), K(k)\boldsymbol{x}_j(k)\right\rangle\right) / Z_{\beta,i}(k)$ is the attention score.

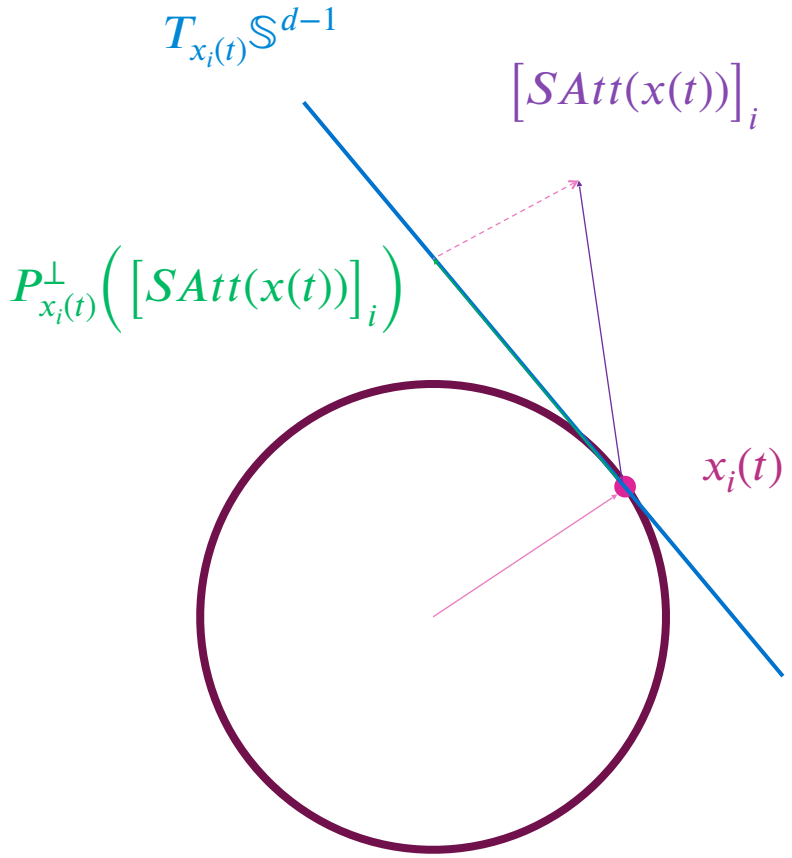
Extension to Transformers

- Transformers operate on a sequence of vectors $\boldsymbol{x}(0) \in \mathbb{R}^{n\times d}$, where each $\boldsymbol{x}_i(0) \in \mathbb{R}^d$ is called a *token/particle*, and $\boldsymbol{x}(0)$ is called a *prompt*.
- Challenge: Transformers use layer normalization

$$\boldsymbol{x}_i(k+1) = \frac{\boldsymbol{x}_i(k) + \left[SAtt\left(\boldsymbol{x}(k); K(k), Q(k), V(k)\right)\right]_i}{\left\|\boldsymbol{x}_i(k) + \left[SAtt\left(\boldsymbol{x}(k); K(k), Q(k), V(k)\right)\right]_i\right\|},$$

- Intuition: A transformer in continuous-time is a flow map on $\left(\mathbb{S}^{d-1}\right)^n$, where $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ is the unit sphere in $\mathbb{R}^d$
  - Project the flow onto the tangent space of the $\mathbb{S}^{d-1}$

$$\dot{x}_i(t) = P^{\perp}_{x_i(t)}\left(\left[SAtt\left(x(t)\right)\right]_i\right), \text{ where } P^{\perp}_x y = y - \langle x, y\rangle x$$



Preliminary Observations in Transformers

- Consider a simple model: $Q(t) = K(t) = V(t) = I$

$$\dot{x}_i(t) = P^{\perp}_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^{n} \exp\left(\beta \left\langle \boldsymbol{x}_i(t), \boldsymbol{x}_j(t) \right\rangle \right) \boldsymbol{x}_j(t) \right)$$

- For most of the paper, we will consider the case $Q(t) = K(t) = V(t) = I$

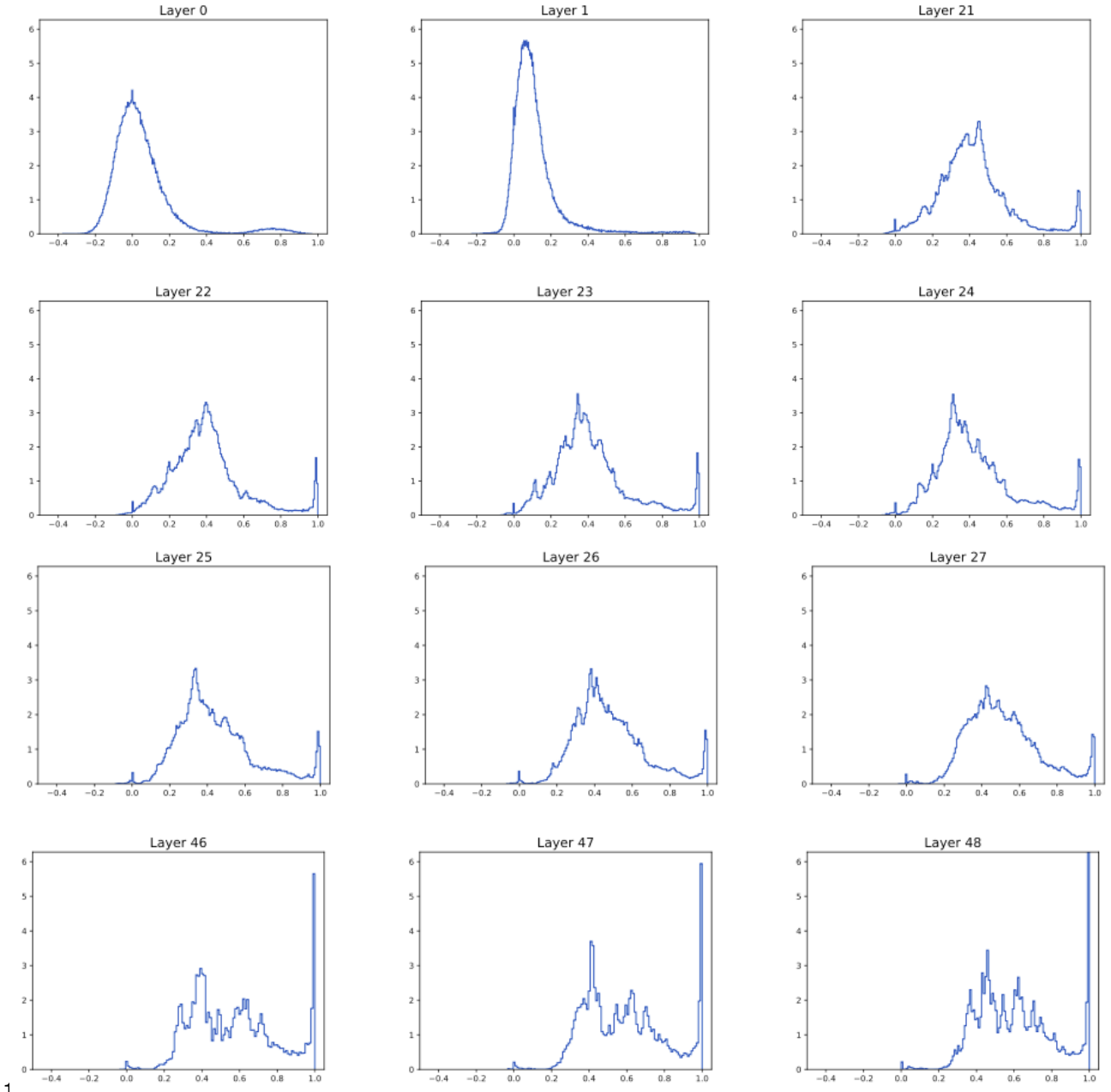


Figure 1. Histogram of $\{\langle x_i(t), x_j(t)\rangle\}_{(i,j)\in[n]^2, i\neq j}$ at different layers t in the context of the trained ALBERT XLarge v2 model ([LCG+20] and https://huggingface.co/albert-xlarge-v2)[3], which has constant parameter matrices. Here we randomly selected a single prompt, which in this context is a paragraph from a random Wikipedia entry, and then generate the histogram of the pairwise inner products. We see the progressive emergence of clusters all the way to the 24th (and last) hidden layer (top), as evidenced by the growing mass at 1. If the number of layers is increased, up to 48 say, the clustering is further enhanced (bottom).

Back to Continuity Equation

- Let $\mu_t(x)$ be the distribution of all the particles $x_1(t), \ldots, x_n(t)$
- Consider the infinite particle limit

$$\dot{x}_i(t) = X[\mu_t](x_i(t))$$

- Where the vector field $X[\mu] : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ is

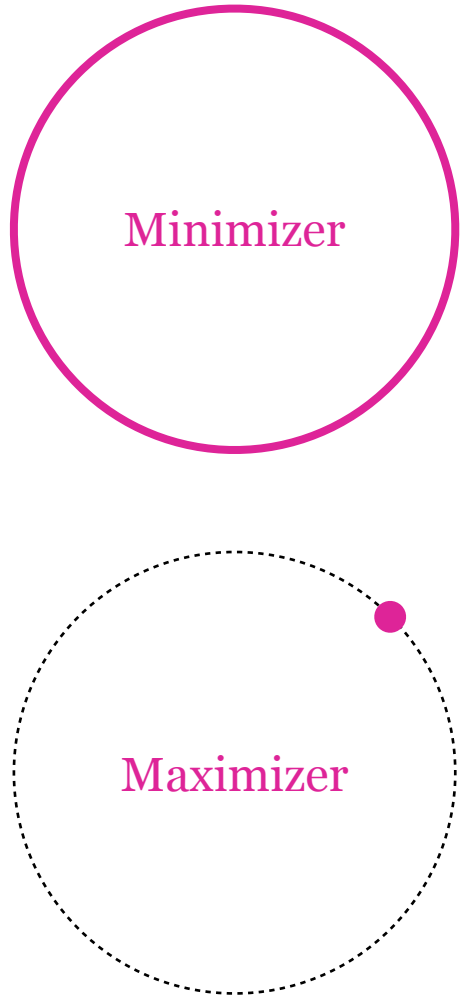
$$X[\mu_t](x) = P_x^{\perp}\left(Z_{\beta,\mu_t}(x)^{-1}\int \exp(\beta\langle x, y\rangle)\ y\ d\mu_t(y)\right)$$

- With $Z_{\beta,\mu_t}(x) = \int \exp(\beta\langle x, y\rangle)\ d\mu_t(y)$
- How does $\mu_t(x)$ evolve with t? Continuity equation!

$$\partial_t \mu_t = -\nabla \cdot (X[\mu_t]\mu_t)$$

The Interaction Energy

- Does the continuity equation admit quantities that are monotonically increasing/ decreasing?
- Answer: Interaction energy
- $E_\beta(\mu_t) = \frac{1}{2\beta}\iint \exp(\beta\langle x, x'\rangle) d\mu_t(x) d\mu_t(x')$
- Its time derivative, $\frac{d}{dt}E_\beta[\mu_t] = \int \|X[\mu_t](x)\|^2 Z_{\beta,\mu_t}(x) d\mu_t(x)$ increases along the continuity equation.
  - Similarly, if $V = -I$, then the interaction energy decreases along the continuity equation.
- **Proposition 3.4.** Let $\beta > 0$ and $d \geq 2$. The unique global minimizer of E_β over $\mathcal{P}(\mathbb{S}^{d-1})$ is the uniform measure. Any global maximizer is a Dirac mass δ_{x^*} at some $x^* \in \mathbb{S}^{d-1}$.
- Proof outline: Consider Gegenbauer polynomials!



Wasserstein Gradient Flow: Attempt #1

- Recall: Suppose a density evolves by:
  $\partial_t p_t = \mp \nabla_{W_2} f(p_t) = \pm \nabla \cdot (p \nabla \nabla_{L_2} f(p))$
- Then a particle drawn from p_t evolves by the ODE
  $\dot{\mathbf{x}}_t = \pm \nabla \nabla_{L_2} f(p_t(\mathbf{x}_t))$
- Notice that $X[\mu]$ is a logarithmic derivative:

$$\nabla_S \log \int \beta^{-1} \exp(\beta \langle x, y \rangle) d\mu(y)$$

$$= P_x^{\perp} \left(Z_{\beta,\mu}^{-1}(x) \int \exp(\beta \langle x, y \rangle) \nabla \langle x, y \rangle d\mu(y) \right)$$

$$= P_x^{\perp} \left(Z_{\beta,\mu}^{-1}(x) \int \exp(\beta \langle x, y \rangle)\ y\ d\mu(y) \right)$$

- However, $\log \int \beta^{-1} \exp(\beta \langle x, y \rangle) d\mu(y)$ cannot be expressed as an L_2 gradient

- Remark: Understanding **Spherical Gradient**
- At a point $x \in \mathbb{S}^{d-1}$, the tangent space is
  $T_x \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \langle v, x \rangle = 0\}$
- Recall gradients:
  $Df(x)[v] = \langle \nabla_S f(x), v \rangle_S$
  $= \langle \nabla f(x), v \rangle - \langle \nabla f(x), x \rangle \langle x, v \rangle$
- Conclusion:
  $\nabla_S f(x) = \nabla f(x) - \langle \nabla f(x), x \rangle x$
  $= P_x^{\perp}(\nabla f(x))$

Attempt #2: Remove denominator!

- Remove $Z_{\beta,i}(t)$ in the softmax formulation (e.g. Sinkformer)
- New attention method:

$$X[\mu_t](x) = P_x^{\perp}\left(\int \exp(\beta\langle x, y\rangle)\ y\ d\mu(y)\right)$$

- Continuity equation:

$$\partial_t \mu_t = -\nabla \cdot \left(P_x^{\perp}\left(\int \exp(\beta\langle x, x'\rangle d\mu_t(x')\mu_t(x)\right)\right)$$

- Back to interaction energy

$$E_\beta(\mu_t) = \frac{1}{2\beta}\iint \exp(\beta\langle x, x'\rangle) d\mu_t(x) d\mu_t(x')$$

- Take the L_2 gradient

$$\nabla_{L_2} E_\beta(\mu_t) = \frac{1}{\beta}\int \exp(\beta\langle x, x'\rangle) d\mu_t(x')$$

- Take the spherical gradient

$$\nabla_S \nabla_{L_2} E_\beta(\mu_t) = P_x^{\perp}\left(\int \exp(\beta\langle x, x'\rangle) x' d\mu_t(x')\right)$$

- Conclusion: $X[\mu](x) = \nabla_S \nabla_{L_2} E_\beta[\mu](x)$
  $\partial_t \mu_t = \nabla_{W_2} E_\beta[\mu] \implies$ interaction energy is increasing with t

Sander at al. "Sinkformers: Transformers with doubly stochastic attention". AISTATS 2022

Sketch: Extension to general parameters

- Assume $Q^T K$ is symmetric and $V = Q^T K$
- New inner product on $T_X(\mathbb{S}^{d-1})^n$:

$$\langle (a_1, \ldots, a_n), (b_1, \ldots, b_n) \rangle_X = \sum_{i=1}^{n} Z_{\beta,i} \langle a_i, b_i \rangle$$

- where $a_i, b_i \in T_{x_i}\mathbb{S}^{d-1}$ and $Z_{\beta,i} = \sum_{j=1}^{n} e^{\beta \langle x_i, x_j \rangle}$

$$\text{Set } E_\beta(X) = \frac{1}{2\beta} \sum_{i=1}^{n} \sum_{j=1}^{n} e^{\beta \langle V x_i, x_j \rangle}$$

- Exercise: Show $\dot{X}(t) = \nabla_X E_\beta(X(t))$, where

$$\dot{X}_i(t) = P_{x_i}^{\perp} \left(\sum_{j=1}^{n} e^{\beta \langle V x_i, x_j \rangle} V x_j \right)$$

CLUSTERING IN TRANSFORMERS

- A single cluster for small/large β
- High-dimensional cases

The case $\beta = 0$ (Theorem 4.1)

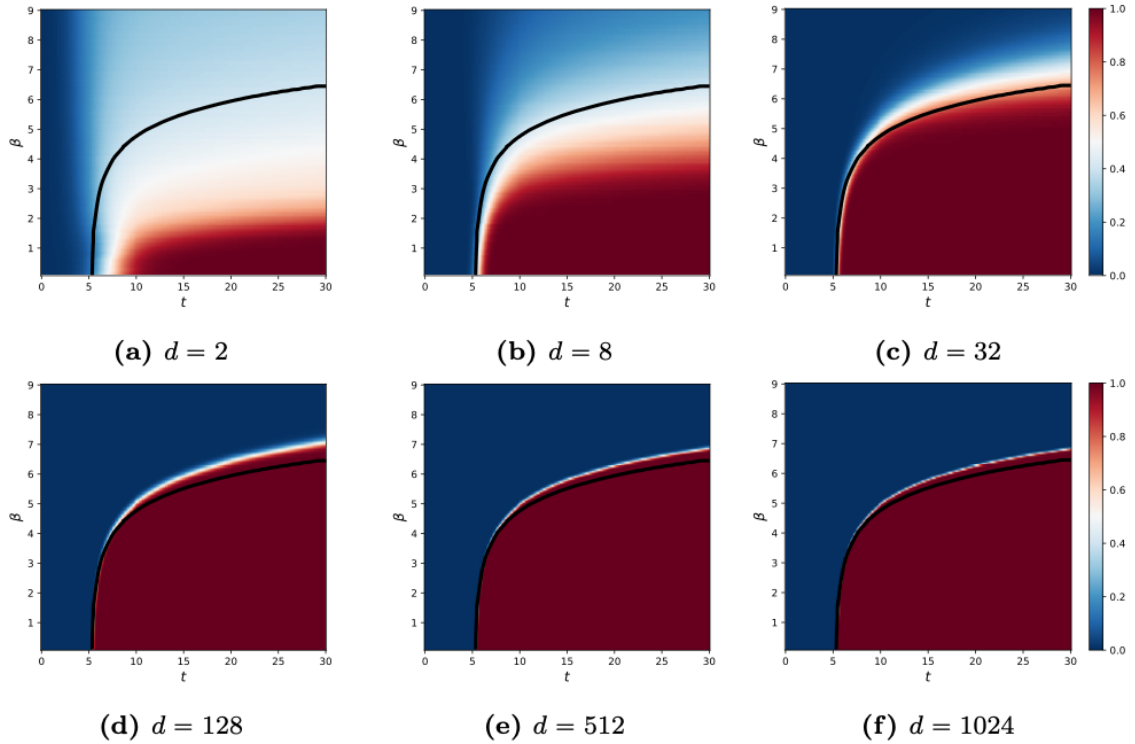
- Vector field: $\dot{x}_t(t) = P^{\perp}_{x_i}(\bar{x}(t))$
- Result: For Lebesgue almost any initial sequence $(x_i(0))_{i\in[n]}$, all particles converge to a single consensus point x^*
- Proof structure:
  - Łojasiewicz theorem (1963): ensures convergence to critical points
  - Benaïm (1999): constructs a Lyapunov function around the saddle manifold

Small and Large β

- Theorem 4.3 (Small β): If $\beta \in O(n^{-1})$, almost all initial sequence $(x_i(0))_{i\in[n]}$ converge to a single cluster.
- Theorem 5.1 (Large β): If $\beta \geq C(d)n^2$, same conclusion holds.
- Proof sketch:
  - Dynamics: Gradient flow of interaction energy.
  - Energy landscape: Global maxima = Dirac masses
  - Łojasiewicz theorem + Benaïm

High Dimensional Results

- Theorem 6.1: In $d \geq 3$, clustering occurs for any $\beta \geq 0$
- Theorem 6.3: If $d \geq n$, convergence is exponential: $\|x_i(t) - x^*\| \leq Ce^{-\lambda t}, \lambda = O(e^{-\beta})$
- Theorem 6.9: If $d \geq n$, histogram of inter-particle inner product collapses to 1 in finite time.



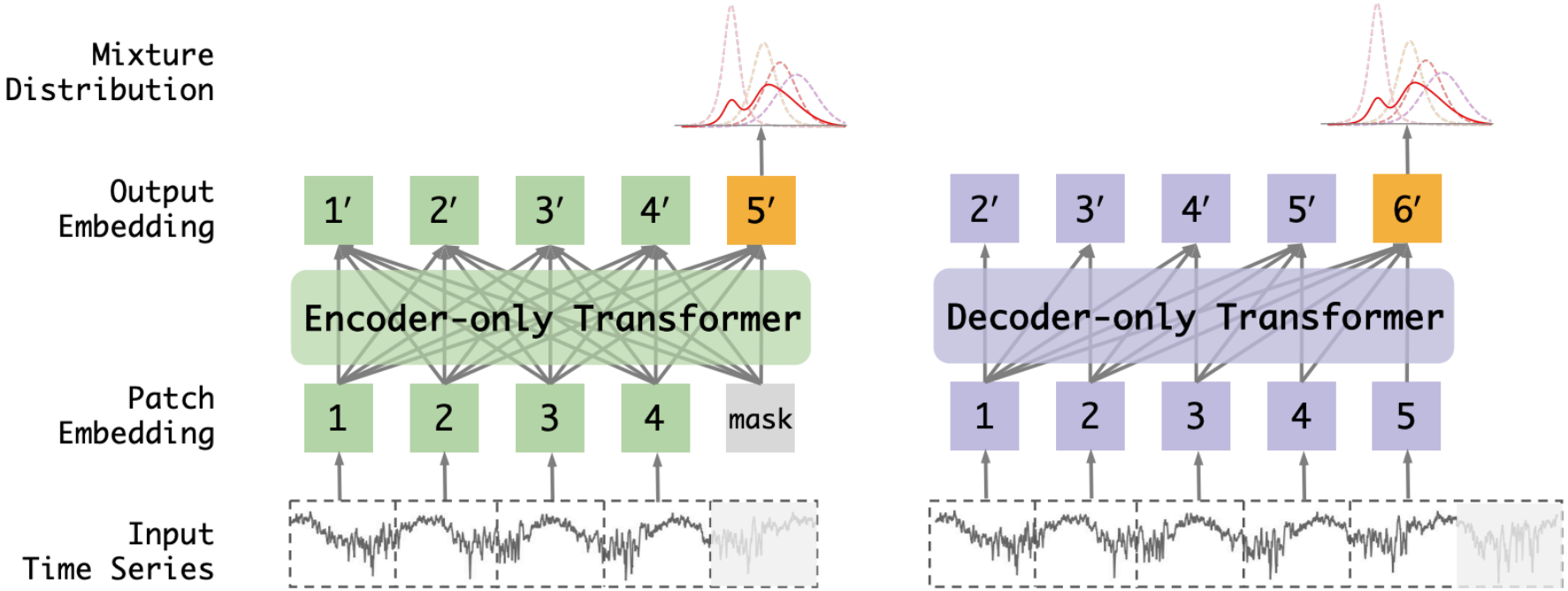
(a) $d = 2$ (b) $d = 8$ (c) $d = 32$

(d) $d = 128$ (e) $d = 512$ (f) $d = 1024$

Figure 3. Plots of the probability that randomly initialized particles following (SA) cluster to a single point as a function of t and β: we graph the function $(t, \beta) \mapsto \mathbb{P}_{(x_1(0),\ldots,x_n(0))\sim\sigma_d}(\{\langle x_1(t), x_2(t)\rangle \geqslant 1 - \delta\})$, which is equal to $(t, \beta) \mapsto \mathbb{P}_{(x_1(0),\ldots,x_n(0))\sim\sigma_d, i\neq j \text{ fixed}}(\{\langle x_1(t), x_2(t)\rangle \geqslant 1 - \delta\})$ by permutation equivariance. We compute this function by generating the average of the histogram of $\{\langle x_i(t), x_j(t)\rangle \geqslant 1 - \delta : (i, j) \in [n]^2, i \neq j\}$ over 2^{10} different realizations of initial sequences. Here, $\delta = 10^{-3}$, $n = 32$, while d varies. We see that the curve $\Gamma_{\infty,\delta}$ defined in (6.12) approximates the actual phase transition with increasing accuracy as d grows, as implied by Theorem 6.9.

Future Research (Maybe course project :))

- Comparing Encoder-only and Decoder-only architectures
- What does the mixture distribution look like?



Yao et al. "Towards neural scaling laws for time series foundation models." ICLR 2025